Document Classification with Reuter50 News Text

Muzhe Zeng
mzeng6@wisc.edu

Abstract

This paper utilizes sever supervised machine learning methods to perform document classification task. The involved machine learning methods are logistic regression, support vector machine, and multi-layer perceptions (deep neural network). The goal is to compare these methods and investigate each methods' strengths in terms of accuracy and speed. The project Github repository is in `https://github.com/MuzheZeng/Document-Classification-with-vsp`.

1 Introduction

Machine learning (ML) is transforming the world's life in all respects. With the growing knowledge in ML area as well as the wide-spread applications, the ML community is expanding at an unprecedented rate. One direct outcome is the production of various ML methodologies. Just like shoppers are usually overwhelmed by the numerous options in the shopping mall and confused about which item is the best to spend money on. ML practitioners are also frequently encountering the question of which ML method to apply or experiment in solving different real-world problems.

ML methods can be categorized into subgroup in many sense: supervised vs un-supervised learning, parametric vs non-parametric, etc. For supervised learning, the most popular ones include but not limit to linear regression (LR) [1], decision tree, support vector machine (SVM) [2], artificial neural network (ANN) [3]. Each method construct the data model from a different perspectives and come with different model assumptions. The computation strategies and the theory behind them are also distinctly diverse. As the ML community grows, is becomes more and more important to gain knowledge about the comparisons between these ML models.

In this paper, we are comparing three popular supervised learning methods: LR, SVM, and ANN. Text classification is presumably hard and is able to reveal most MLs' weaknesses. Therefore the comparison will be focused on high-dimensional unstructured text dataset and the task document classification.

After the introduction section, we will describe the dataset in Section 2. Then we revisit the basics of these three popular ML methods in Section 3. Section 4 summarizes the experiment result and we conclude our paper in Section 5.

2 Dataset

The dataset is collected from UCI: `https://archive.ics.uci.edu/ml/datasets/Reuter_50_50`. The dataset is split in halves for training and testing purpose, with 2500 documents each. The response variable is the author of each text, and there are 50 authors (then 50 categorical levels). The response variable is perfectly balanced in both train and test set.

2020 Fall, Matrix Methods in Machine Learning (CS 532) Final Project.

An example of a instance in our training set is displayed below:

> Drugstore giant Revco D.S. Inc. said Monday it agreed to buy regional chain Big B Inc. in a sweetened takeover valued at $380 million. The transaction calls for Twinsburg, Ohio-based Revco to buy all outstanding shares of Big B common stock for $17.25 per share, up from Revco's unsolicited offer of $15 per share, which Big B rejected last month. "We are very excited about the combination of Revco and Big B. I am pleased we were able to bring this process to a fast and successful conclusion," said Dwayne Hoven, president and chief executive officer of Revco. The deal will ...

The example above has 677 words and belongs to the writer RobinSidel. The average length of the train set is 488 words and 495 for test set. The names of all the writers (categorical labels) are listed below.

```
['AaronPressman' 'AlanCrosby' 'AlexanderSmith' 'BenjaminKangLim'
 'BernardHickey' 'BradDorfman' 'DarrenSchuettler' 'DavidLawder'
 'EdnaFernandes' 'EricAuchard' 'FumikoFujisaki' 'GrahamEarnshaw'
 'HeatherScoffield' 'JanLopatka' 'JaneMacartney' 'JimGilchrist'
 'JoWinterbottom' 'JoeOrtiz' 'JohnMastrini' 'JonathanBirt' 'KarlPenhaul'
 'KeithWeir' 'KevinDrawbaugh' 'KevinMorrison' 'KirstinRidley'
 'KouroshKarimkhany' 'LydiaZajc' "LynneO'Donnell" 'LynnleyBrowning'
 'MarcelMichelson' 'MarkBendeich' 'MartinWolk' 'MatthewBunce'
 'MichaelConnor' 'MureDickie' 'NickLouth' 'PatriciaCommins'
 'PeterHumphrey' 'PierreTran' 'RobinSidel' 'RogerFillion' 'SamuelPerry'
 'SarahDavison' 'ScottHillis' 'SimonCowell' 'TanEeLyn' 'TheresePoletti'
 'TimFarrand' 'ToddNissen' 'WilliamKazer']
```

The most frequent vocabularies in the train set is:

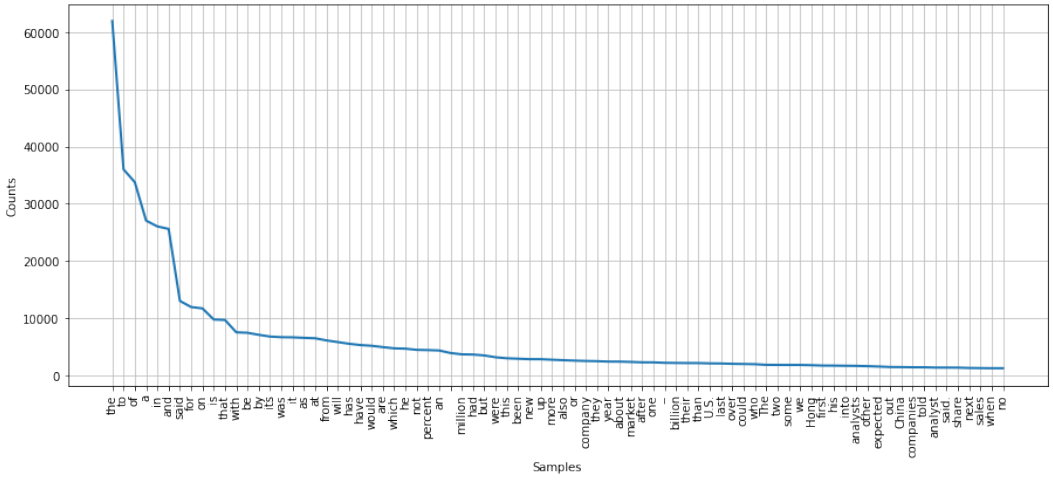

Figure 1: Top 80 words appear in the train set. Most of them are stop-words. Our ML experiment will remove these non-informative signals.

3 Methods

3.1 Logistic Regression

For binary logistic regression, the response variable is a two-level categorical column. The generalized linear model language suggests the model relation as

$$z = X\theta = \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p. \tag{1}$$

and the probability of response variable y being positive $(= 1)$ is

$$\Pr(y = 1) = g(z) = \frac{e^z}{1 + e^z}. \tag{2}$$

The right hand side of Equation (2) is called link function.

3.2 Support Vector Machine

SVM is a versatile and powerful supervised ML, which is able to deal with linear or nonlinear classification, regression problems. For regression purpose, SVM aims to fit the largest possible "streets" such that most samples fall within the area between the "streets". The "kernel trick" in the SVM theory facilitates the computation for feature extensions. This is a key step for SVM to handle nonlinearity in the data. Four popular kernel options are:

Linear $K(a, b) = a^T b)$.

Polynomial $K(a, b) = (\gamma a^T b + r)^d$.

RBF $K(a, b) = \exp(-\gamma \|a - b\|^2)$.

Sigmoid $K(a, b) = tanh(\gamma a^T b + r)$.

3.3 Artificial Neural Network

Artificial Neural Network (ANN) is the foundation of DL. The model architecture is inspired by the computational neurobiology domain to mimic the signal transportation mechanism among neurons and synapses in the brain. The further development of Back-Propagation (BP) algorithm and automatic differentiation, stochastic gradient descent (SGD), batch normalization, GPR integration, etc, all enable the wide spread application of DL.

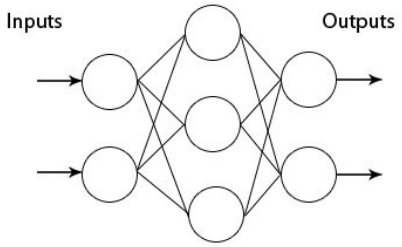

Figure 2: A visualization of 3-layer artificial neural network.

4 Experiment

Logistic regression does not require hyper-parameter tuning step. For SVM, there are: constraint factor (C), kernel option. For ANN, there are: hidden layer size, number of layers, activation functions, penalty parameter (α), etc. We tune the hyper-parameters by grid search cross validation using only the training set. Eventually, we select the following hyper-parameters:

SVM $C = 10$, kernel = sigmoid.

ANN Hidden layer structure = (100, 100), activation function = logistic, $\alpha = 0.1$.

Model	CV accuracy	test accuracy	training time (seconds)
Linear Regression	0.85	0.75	55
Support Vector Machine	0.83	0.67	330
Artificial Neural Network	0.87	0.76	297

Table 1: Experiment Results. Comparisons of accuracy and speed.

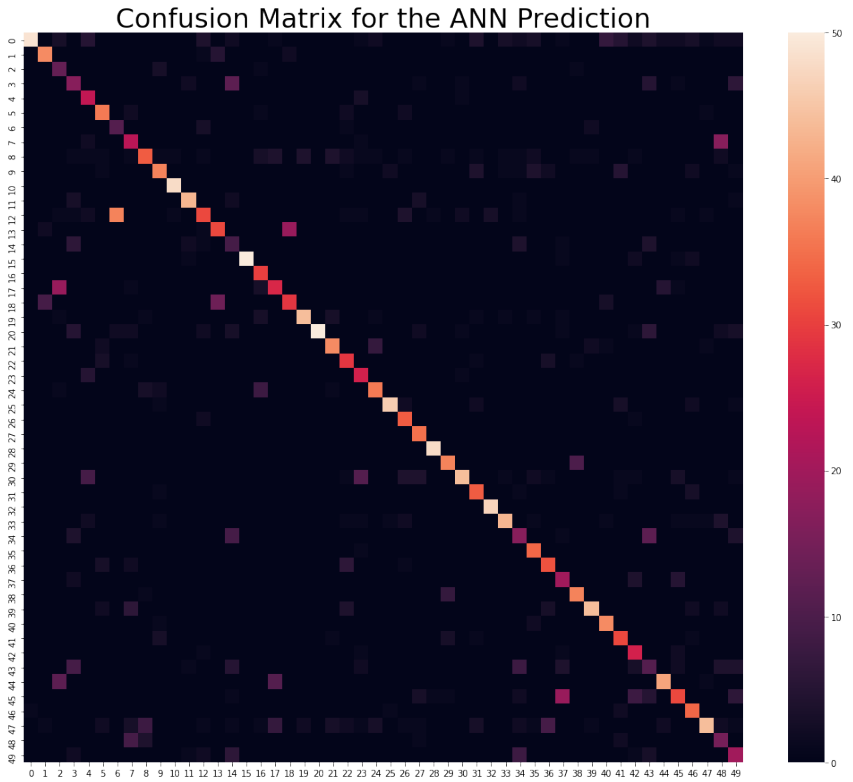

Figure 3: Confusion matrix for the best model over the test set. The lighter-colored tiles have more samples. Off-diagonal tiles are all mis-classified samples.

Table 1 summarizes the result of the experiment. The winner among these three candidate ML methods is the famous ANN. With accuracy in the test set as high as 76%, which is much better than random guess (2%). Machine learning is popular for a reason! LR is slightly worse than ANN in terms of accuracy. SVM performs the worst among the three. In terms of latency, LR is much faster than the other two. If the classification task is designed for a product in which latency is a major concern, then LR will be the best option among the three. If one only cares about accuracy, then ANN is the choice.

Figure 3 demonstrates the confusion matrix of the 50 classes. From this matrix we can investigate which authors' writings are easily mis-recognized. Firstly, most weights fall in the diagonal entries, indicating accurate predictions. By looking at the bright off-diagonal tiles, we can see that ANN sometimes mistaken David Lawder as Heather Scoffield, Alexander Smith as Joe Ortiz, Peter Humphrey as Therese Poletti.

5 Discussions

This paper compares three popular machine learning methods in document classification task. The Artificial Neural Network is the most accurate while Logistic Regression is the fastest. With the help of GPU and parallel computing techniques, one can actually get the best of most worlds. However, the difference between cross-validation error rate and the test set prediction error rate implies certain level of over-fitting. In order to improve the models, one might need to expand the grid search in

the hyper-parameter spaces. Also, some feature engineering prior work might further strengthen the signal and therefore improve the performance.

References

[1] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[2] Cortes, Corinna, and Vladimir Vapnik. "Support vector machine." Machine learning 20, no. 3 (1995): 273-297.

[3] Hassoun, Mohamad H. Fundamentals of artificial neural networks. MIT press, 1995.